

Log Activity

IF 4090 – Kerja Praktek

Minggu : 4

Perusahaan / Organisasi : *National Institute of Technology, Gifu College* (NIT – GC)

Pembimbing Kerja Praktek : Dr. Eng. Ayu Purwarianti, ST., MT.



Nama : Gerald Dzakwan

NIM : 13514065

Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

2017

Log Activity – Minggu ke-4

Hari, Tanggal	Kegiatan	Hasil
Senin, 3 Juli 2017	<p>1. Mendefinisikan <i>feature</i> yang akan dilatih pada algoritma CRF.</p> <p>2. Membuat <i>script</i> untuk melakukan <i>training</i> model dengan <i>tools</i> sklearn-crfsuite.</p>	<p>1. Fungsi yang mengekstraksi <i>feature</i> dari <i>input message</i>.</p> <p>2. Python <i>script</i> yang dapat melatih model dan menyimpan model.</p>
Selasa, 4 Juli 2017	<p>1. Eksplorasi cara untuk menentukan kedekatan informasi antara dua kata. Dua kata yang dibandingkan yakni kata-kata dalam <i>social media message</i> dan kata-kata dalam <i>user profile</i>.</p> <p>2. Ekplorasi beberapa <i>dataset</i> yang mungkin. Diputuskan untuk implementasi menggunakan Google. Jumlah kemunculan (<i>occurrence</i>) dalam hal ini yakni jumlah halaman web yang muncul saat melakukan pencarian dengan dua kata tersebut.</p>	<p>1. Cara yang dipilih yakni menggunakan <i>co-occurrence metrics</i> dari <i>huge dataset</i>.</p> <p>2. Python <i>script</i> yang memanfaatkan Google Custom Search API untuk menghitung <i>co-occurrence metrics</i> antara dua kata.</p>
Rabu, 5 Juli 2017	<p>1. Melanjutkan CRF <i>algorithm script</i>, membenarkan <i>bug-bug</i> yang ditemukan, dan melakukan modularisasi program.</p> <p>2. Mengeksplorasi WordNet untuk mengukur <i>sentence similarity</i> antara <i>original message</i> dan <i>anonymized message</i>.</p>	<p>1. <i>Script</i> telah dapat memprediksi <i>named entity</i> dengan format yang benar.</p> <p>2. Python <i>script</i> untuk mengukur <i>sentence similarity</i>, namun masih <i>buggy</i>.</p>
Kamis, 6 Juli 2017	<p>1. Menyusun metode dalam bentuk <i>diagram workflow</i> untuk menentukan frase mana yang benar-benar <i>private</i>. Secara umum, dilakukan dalam dua tahap, yakni <i>co-occurrence metrics</i> (telah dikerjakan) dan <i>rule-based approached</i> (belum dikerjakan). <i>Workflow</i> ini nantinya akan dibahas bersama Prof. Yasuda pada <i>review</i> hari Jumat.</p>	<p>1. Diagram <i>workflow</i> / cara kerja program dalam mengidentifikasi beberapa <i>private phrase</i> seperti <i>person</i>, <i>location</i>, <i>organization</i>, dan <i>time</i>.</p>

<p>Jumat, 7 Juli 2017</p>	<ol style="list-style-type: none"> 1. 4th Weekly Report dengan Prof. Yasuda. 2. Memulai pengerjaan identifikasi <i>private phrase</i> dengan <i>rule-based approach</i>. 	<ol style="list-style-type: none"> 1. Progress seminggu terakhir tersampaikan ke Prof. Yasuda. Mendapatkan <i>feedback</i> bahwa untuk <i>threshold</i> dari <i>co-occurrence metrics</i> sebaiknya didefinisikan dalam <i>range</i>. Metode yang diajukan untuk mengurangi jumlah <i>replacement</i> sudah baik. Namun, lebih bagus lagi jika dirancang sebuah mekanisme untuk menentukan apakah informasi yang tersisa dalam teks dapat digunakan untuk mengidentifikasi seseorang. 2. Fungsi-fungsi Python untuk mengekstrak <i>candidate phrase</i> dan menyeleksi <i>private phrase</i>.
-------------------------------	---	---

Log Activity – Minggu ke-3

Hari, Tanggal	Kegiatan	Hasil
Senin, 26 Juni 2017	<ol style="list-style-type: none"> 1. Membuat <i>postprocessing module</i> untuk meng-<i>construct</i> kembali kalimat dari struktur <i>nlk.tree.Tree</i> ke struktur kalimat aslinya. 2. Membuat <i>simple replacer</i>, yakni mengganti <i>named entity</i> dengan kelasnya masing-masing. 3. Menambahkan mekanisme <i>coreference resolution</i>, yakni mengidentifikasi seluruh entitas yang sama dan menggantinya dengan frase yang sama pula. 	<ol style="list-style-type: none"> 1. Penambahan modul baru, yakni <i>postprocessing module</i> pada sistem. 2. Penambahan fungsionalitas baru yakni penggantian <i>named entity</i> pada modul <i>anonymization</i>. 3. Penambahan fungsionalitas baru yakni <i>coreference resolution</i> pada modul <i>anonymization</i>.
Selasa, 27 Juni 2017	<ol style="list-style-type: none"> 1. Melakukan evaluasi dari <i>classifier</i> yang sudah dibuat dan menganalisis penyebab tingkat akurasi yang rendah dari <i>classifier</i>. 2. Mengeksplorasi algoritma-algoritma baru yang meng-<i>consider feature</i> dari kata-kata pada satu kalimat. Setelah dipertimbangkan, saya memilih untuk mengeksplorasi lebih jauh algoritma <i>Conditional Random Field (CRF)</i>. 	<ol style="list-style-type: none"> 1. Mengetahui penyebab rendahnya akurasi, yakni <i>corpus dataset</i> yang terbatas dan algoritma <i>training model</i> yang hanya meng-<i>consider feature</i> dari kata yang akan diprediksi saja. 2. Mengetahui prinsip kerja dari algoritma CRF, khususnya <i>linear chain CRF</i>.
Rabu, 28 Juni 2017	<ol style="list-style-type: none"> 1. Mengeksplorasi <i>tools</i> yang dapat digunakan untuk mengimplementasi algoritma <i>linear chain CRF</i>. 2. Melakukan migrasi <i>corpus</i> dari <i>text file</i> ke satu basis data yang terintegrasi agar mudah untuk menambahkan <i>corpus</i> baru (agar <i>scalable</i>). 	<ol style="list-style-type: none"> 1. Ditentukan <i>tools</i> yang dapat digunakan, yakni <i>sklearn-crfsuite</i> versi 0.3. 2. Basis data MySQL berisi data seluruh <i>corpus</i> yang digunakan. Basis data menyimpan kata (<i>word</i>), <i>pos tag</i> dari kata, dan <i>named entity</i> dari kata.
Kamis,	1. 3 rd Weekly Report dengan Prof. Yasuda	1. Progress seminggu terakhir

29 Juni 2017		tersampaikan ke Prof. Yasuda dan mendapatkan <i>feedback</i> bahwa perlu ada mekanisme dan kriteria untuk mengukur seberapa jauh <i>message</i> bisa dikatakan sudah <i>private</i> . Harapannya, program memiliki kinerja bagus yakni dapat sesedikit mungkin melakukan <i>replacement</i> sehingga informasi masih <i>useful</i> dan <i>distance</i> antara teks asli dan <i>anonymized text</i> sedekat mungkin.
Jumat, 30 Juni 2017	<ol style="list-style-type: none"> 1. Melanjutkan proses migrasi <i>corpus</i> ke <i>database</i> (membutuhkan proses yang lama dan perlu <i>preprocessing script</i> yang berbeda-beda untuk setiap <i>corpus</i> karena struktur setiap <i>corpus</i> berbeda pula). 2. Melakukan eksplorasi penggunaan <i>tools</i> <i>sklearn-crfsuite</i> 0.3. 	<ol style="list-style-type: none"> 1. Basis data MySQL berisi data seluruh <i>corpus</i> yang digunakan. Basis data menyimpan kata (<i>word</i>), <i>pos tag</i> dari kata, dan <i>named entity</i> dari kata. 2. Mengetahui cara implementasi pendefinisian <i>feature</i> dan cara melakukan <i>training</i> model pada <i>sklearn-crfsuite</i>.

Log Activity – Minggu ke-2

Hari, Tanggal	Kegiatan	Hasil
Senin, 19 Juni 2017	1. Mempelajari struktur data dari <i>corpus</i> yang digunakan. 2. Mempelajari <i>scikit-learn module</i> melalui buku “ <i>Hands-On Machine Learning with Scikit-Learn & TensorFlow</i> ” karangan Aurelion Geron Penerbit O’Reilly.	1. Mengetahui bagaimana cara menggunakan dan memanfaatkan <i>corpus</i> tersebut. 2. Mengetahui cara membuat <i>classifier</i> , melakukan <i>training</i> dan <i>testing</i> menggunakan <i>scikit-learn</i> .
Selasa, 20 Juni 2017	1. Mencoba membuat beberapa <i>classifier</i> seperti Naive Bayes dan Perceptron.	1. <i>Classifier</i> program untuk <i>training corpus</i> .
Rabu, 21 Juni 2017	1. Membuat dokumen <i>term of reference</i> (TOR). 2. Melakukan <i>debugging</i> terhadap program <i>classifier</i> .	1. Dokumen TOR dalam bentuk <i>hardcopy</i> . 2. <i>Bug fixed</i> .
Kamis, 22 Juni 2017	1. 2 nd Weekly Report dengan Prof. Yasuda dan <i>review</i> TOR. 2. Melakukan eksperimen <i>training corpus</i> dengan <i>classifier</i> dan jumlah <i>sample data train</i> yang berbeda-beda.	1. Progress seminggu terakhir tersampaikan ke Prof. Yasuda dan mendapatkan <i>feedback</i> untuk mencoba metode <i>k-anonymization</i> . Namun, setelah dieksplorasi, metode tersebut tidak sesuai karena membutuhkan <i>structured data</i> sebagai input (semisal <i>record-record</i> pada <i>database</i>). Sedangkan, <i>research project</i> saya menggunakan <i>unstructured data</i> (<i>message</i>) sebagai input. 1. Model hasil <i>training</i> .
Jumat, 23 Juni 2017	1. Melakukan <i>debugging</i> kembali terkait masalah-masalah seperti <i>MemoryError</i> . 2. Membuat mekanisme <i>save/load file</i> model dan melakukan <i>testing</i> terhadap sejumlah <i>input</i>	1. <i>Bug fixed</i> dengan menggunakan <i>partial fit</i> . 2. Program <i>classifier</i> termodifikasi untuk <i>save/load</i>

	<i>message.</i>	model. Mengetahui bahwa <i>classifier</i> masih belum akurat dalam mendeteksi kelas yang tepat untuk sebuah <i>named-entity</i> .
--	-----------------	---

Log Activity – Minggu ke-1

Hari, Tanggal	Kegiatan	Hasil
Senin, 12 Juni 2017	1. Tiba di <i>Kansai International Airport</i> , Jepang. 2. <i>Welcome ceremony</i> dengan Mr. Yoshito Itoh (<i>School Principal</i>) di NIT – GC dan pemberian uang beasiswa untuk tiga minggu pertama.	1. - 2. -
Selasa, 13 Juni 2017	1. <i>Research Introduction</i> oleh Profesor Tajima (<i>Computer Network Lab</i>) dan Profesor Yasuda (<i>Artificial Intelligence Lab</i>). 2. Perkenalan diri dengan para siswa yang berada di Prof. Tajima dan Prof. Yasuda Lab.	1. Topik riset terpilih, yakni <i>Message Anonymization using Trained Named Entity Recognition</i> . Riset dilakukan di Prof. Yasuda Lab. 2. -
Rabu, 14 Juni 2017	1. Mempelajari <i>paper</i> berjudul “ <i>Anonymizing Private Phrases and Detecting Disclosure in Online Social Networks</i> ” yang merupakan disertasi dari Prof. Nguyen Son Hoang Quoc. 2. Kegiatan mingguan <i>talk cafe</i> , yakni <i>foreign students</i> berinteraksi dengan siswa NIT – GC.	1. Mengetahui variasi metode yang dapat digunakan untuk mendeteksi <i>named-entity</i> dan untuk <i>me-replace named-entity</i> . 2. -
Kamis, 15 Juni 2017	1. Mempelajari <i>tools</i> dan <i>Python libraries</i> yang akan digunakan, yakni <i>Natural Language Toolkit</i> dari buku karangan Steven Bird, Ewan Klein, dan Edward Loper Penerbit O’Reilly. 2. <i>1st Weekly Report</i> dengan Prof. Yasuda. 3. <i>Welcome party</i> dengan siswa NIT – GC.	1. Mengetahui cara melakukan <i>preprocessing</i> dari <i>raw text</i> , cara melakukan <i>NP chunking</i> , dan cara melakukan <i>named entity recognition</i> . 2. Progress seminggu terakhir tersampaikan ke Prof. Yasuda dan mendapatkan <i>feedback</i> . 3. -
Jumat, 16 Juni 2017	1. Mulai membuat program sederhana dalam bahasa Python untuk mendeteksi <i>named-entity</i> dengan <i>natural language toolkit library</i> .	1. <i>Source code</i> program dalam Python yang dapat dilihat di https://github.com/geraldzakwan/TextAnonymizationGNCT .